Augmenting Language Models with a Memory Structure for Improving Consistency of Model Beliefs

Kushal Jain (2019111001) and Abhijit Manatkar (2019101108)

1 Introduction

Large language models contain an enormous corpus of world knowledge, yet when queried, they still generate incoherent replies, even after specialized training. As a result, it becomes difficult to determine what the model actually "believes" about the environment, rendering it prone to inconsistent behavior and simple mistakes. In general, a system is said to (appear to) believe a proposition p, such as "eagles are birds," if it produces responses consistent with p (and its other beliefs). In this project we explore different techniques to augment LMs with an external memory of beliefs and associated mechanisms which help build up an overall system with consistent beliefs.

2 Belief Bank

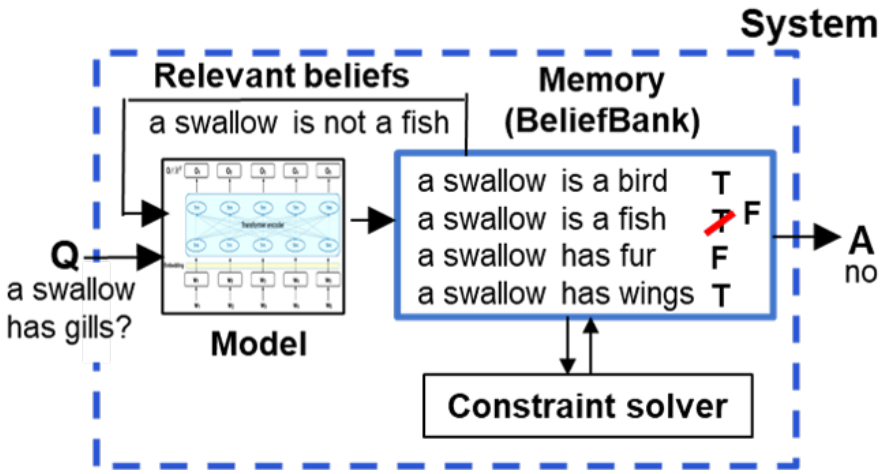
"BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief" [1] introduced a novel method for augmenting an LM with an external memory of beliefs. In this work, the LM is embedded in a system which also contains a store of symbolic memory called the Belief Bank. The contents of the Belief Bank are triples consisting of a sentence S, a truth value T (True/False), and a weight w. This triple indicates that the system believes that sentence S is True/False as indicated by T with a weight w. These "beliefs" are accumulated by posing sentences (S's) as questions to the model and making the model choose between True/False output. The model's confidence in the True/False output is taken as the weight of its belief. This method describes two mechanisms involving its Belief Bank:

1. **Constraint Solving** - The Belief Bank, prior to the experiment is initialized with certain constraints about beliefs. A constraint is an implication of the type $((S_1.T_1 \rightarrow S_2.T_2), w)$ which means that sentence S_1 having truth value T_1 implies sentence S_2 has truth value T_2 and the system violating this constraint will result in it receiving a penalty w. These weights along with the weights of pre-existing beliefs in the Belief Bank

and the confidence of the output to an answer are passed as input to a maximum constraint satisfiability solver (MaxSAT), based on which, it is decided what belief should be added to the Belief Bank or what beliefs in the Belief Bank should have its truth value flipped in order to maintain maximum consistency among beliefs.

2. **Feedback** - When presented with a question, the system searches its Belief Bank and finds beliefs that are relevant to the question asked and adds them as context to the question. The hypothesis is that this helps the model pay "attention" to the different aspects highlighted by the beliefs already known by it (i.e. the one's in the Belief Bank) so that while answering the current question, it can give an answer that is consistent with its existing beliefs.

Figure 1: Diagram of the BeliefBank system [1]



It is observed in the experiments conducted in this study that with exposure to more knowledge, this system is able to build a largely consistent system of beliefs which it can leverage to tune its answers to questions, showing a greatly improved consistency over simple baselines.

3 Project Scope

1. Implement the original Belief Bank, including the constraint-solving and feedback mechanisms.

2. Conduct experimental evaluation and ablation studies as described in the original paper and report results.

3. Propose new approaches to integrating a system of beliefs with a Language Model along with hypotheses about their potential advantages. Conduct experiments to test out the hypotheses. New approaches for introducing belief systems will be around the lines of:

    - Modifying the belief bank store structure (We could consider a key-value store with the key being some sort of learned, pre-existing representation.)
    - Modifying the constraint solving mechanism (One potential case study could be that of examining the effect of restricting constraint solving only to a small subset of constraints/only constraints closely related to the topic.)
    - Modifying the feedback mechanism (Introducing more complex feedback selection strategies)
    - Introducing new mechanisms and proposing advantages.

4 Other Related Work

4.1 Language Models as Knowledge Bases

Recent advancements in language model pretraining on huge textual corpora resulted in a wave of gains for downstream NLP jobs. While acquiring linguistic knowledge , these models may be storing relational knowledge from the training data, and they may be able to answer questions phrased as "fill-in-the-blank" cloze statements. Language models, as demonstrated in [2], offer several benefits over structured knowledge bases, including the fact that they do not require schema engineering, allow practitioners to query about an open class of relations, are straightforward to extend to new data, and do not require human supervision to train [2]. Hence they can be embedded in a broader system that contains an evolving memory of beliefs.

4.2 Logic Framework for Consistency

The incoherence of large language models can be formalized as a generalization of prediction error . [3] presents a learning framework for limiting models using logic rules to keep them consistent.

4.3 Augmenting Language Models with Memory

RAG [3] and REALM [4] are two recent methods which augment language models with a memory structure, but they do it in the form of an external memory that can be used for retrievals either during pre-training or inference. In our case, we build up the memory from the model outputs and incorporate mechanisms which modify this memory to maximize for consistency.

5 Available Datasets and Models

5.1 Datasets

The Belief Bank dataset is openly available at `https://allenai.org/data/beliefbank` The original paper uses Macaw, a T5 based question answering model as the black-box Language Model in its implementation. It is openly available at `https://huggingface.co/allenai/macaw-answer-11b` .
An existing implementation of the Belief Bank system is not openly available so implementing that will be part of the project scope.

5.2 Evaluation of Models

Like the original paper, we will evaluate the consistency and accuracy of beliefs accumulated in the resultant Belief Bank of each model. Accuracy will be measured using the F1 score calculated with reference to ground truth beliefs. Consistency is measured using the metric defined in [3].

6 Tentative Schedule

Task	Deadline
Implementation of constraint solving mechanism	22/10/2022
Implementation of feedback mechanism	22/10/2022
Implementation of Belief Bank and experiments (including ablation studies)	29/10/2022
Implementing new approaches and running experiments	15/11/2022

References

[1] Kassner, Nora, et al. "BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief." arXiv preprint arXiv:2109.14723 (2021).

[2] F. Petroni, Tim Rocktäschel, Patrick Lewis, A. Bakhtin, Yuxiang Wu, Alexander H. Miller, and S. Riedel. 2019. Language models as knowledge bases? In EMNLP.

[3] Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In EMNLP.

[4] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, V. Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval augmented generation for knowledge-intensive nlp tasks. In NeurIPS.